Q8

Understand the working of linear and non-linear regression models. Give a pseudocode and illustrate the same over a sample dataset of your choice.

1. Linear Regression

Introduction:

Linear Regression is a statistical approach to modelling the relationship between an input and output as a linear relationship.

Given a data set of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modelled through a disturbance term or error variable *ε -* an unobserved random variable that adds ‘noise’ to the linear relationship between the dependent variable and regressors. Thus, the model takes the form,

Where T denotes the transpose, so that is the inner product between vectors and .

Pseudo Code:

1. Read Number of Data (n)
2. For i=1 to n:

Read Xi and Yi

Next i

1. Initialize:

sumX = 0

sumX2 = 0

sumY = 0

sumXY = 0

1. Calculate Required Sum

For i=1 to n:

sumX = sumX + Xi

sumX2 = sumX2 + Xi \* Xi

sumY = sumY + Yi

sumXY = sumXY + Xi \* Yi

Next i

1. Calculate Required Constant a and b of y = a + bx:

b = (n \* sumXY - sumX \* sumY)/(n\*sumX2 - sumX \* sumX)

a = (sumY - b\*sumX)/n

1. Display value of a and b

Trace:

For sample data,

Suppose actual line parameters are

Y = mX + c

Where m = slope = 5

C = intercept = 10

N = 5

|  |  |
| --- | --- |
| X | Y |
| 1 | 16 |
| 2 | 19 |
| 3 | 26 |
| 4 | 32 |
| 5 | 36 |

Now, we calculate sums,

sumX = 1 + 2 + 3 + 4 + 5 = 15

sumY = 16 + 19 + 26 + 32 + 36 = 129

sumX2 = 1 + 4 + 9 + 16 + 25 = 55

sumXY = 16 + 38 + 78 + 128 + 180 = 440

Now, we find parameters,

m = (N \* sumXY - sumX \* sumY)/(N\*sumX2 - sumX \* sumX) = (5\*440 – 15\*129) / (5\*55 – 15\*15)

= 5.3

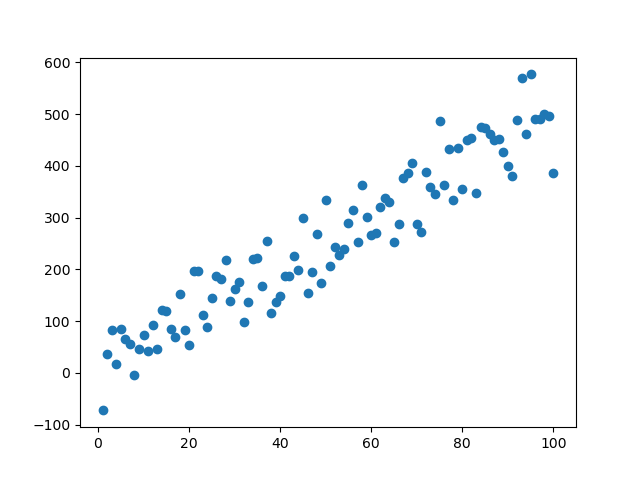
c = (sumY - m\*sumX)/N = (129 – 5.3\*15) / 5 = 9.9

So, Evaluation,

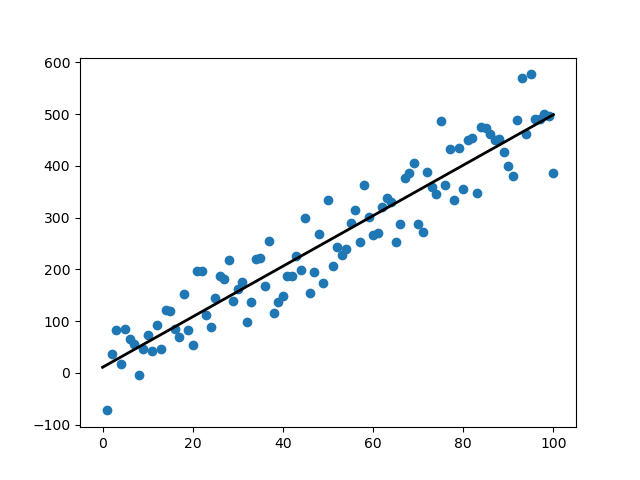
|  |  |  |
| --- | --- | --- |
|  | Actual | Model |
| Slope | 5 | 5.3 |
| Intercept | 10 | 9.9 |

Larger Dataset,

Input Data



Output Model



Parameters:

Slope: Actual: 5 - Predicted: 4.880934505955754

Intercept: Actual: 10 - Predicted: 10.804514757031138

Predictions:

X: 12.3 Actual Y: 71.5 - Predicted Y: 70.84000918028691

X: 10.1 Actual Y: 60.5 - Predicted Y: 60.10195326718424

X: 25.2 Actual Y: 136.0 - Predicted Y: 133.80406430711614

1. Logistic Regression

Introduction:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model (a form of binary regression).

Pseudo Code:

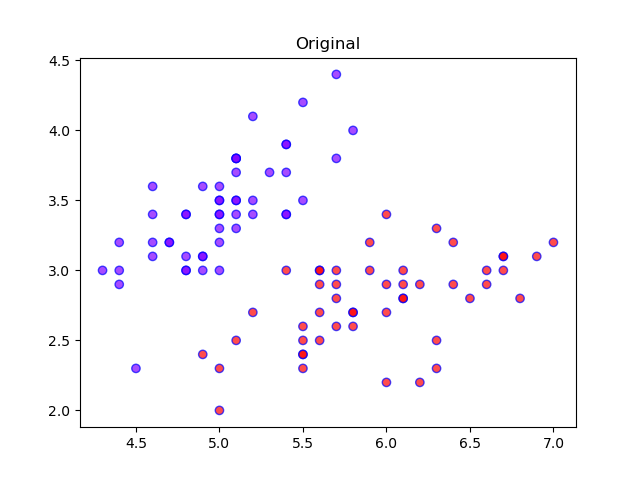
Given Dataset, first we fit the logistic regression model to the train data.

1. Initialize the parameters
2. Repeat {  
    Make a prediction on y  
    Calculate cost function  
    Get gradient for cost function  
    Update parameters  
   }

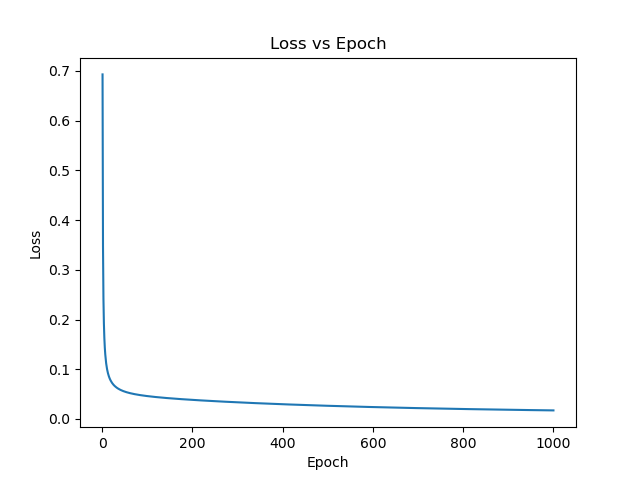
After training, we can use the model parameters to predict for new data points

Trace:

Input Data



Training/Fitting



Output Predictions

